自然语言处理技术赋能 AIGC 识别研究进展*

王伟正[1] 乔鸿[2] 李肖俊[3,4] 王静静[5]

（1 山东师范大学图书馆，山东济南 250358；2 山东师范大学商学院，山东济南 250358；3 齐鲁工业大学（山东省科学院）数字人文研究中心，山东济南 250014；4 齐鲁工业大学（山东省科学院）情报研究所，山东济南 250014；5 山东大学新闻与传播学院，山东济南 250100）

摘要：[目的/意义]随着大语言模型的快速崛起，AIGC 在我们日常生活中无处不在。为防止 AIGC 滥用，减少虚假消息、学术不端、欺骗评论等问题的产生，对自然语言处理技术赋能 AIGC 识别研究进展进行归纳与展望。[方法/过程]首先，明确 AIGC 识别是二值分类问题，其目标是识别一段内容是否是由人工智能生成。然后，采用系统综述方法梳理了 AIGC 识别领域的主要研究成果。[结果/结论] 研究发现全面的优秀数据集对构建 AIGC 识别分类器的重要性，同时探究了当前流行数据集的局限性和发展目标，以及潜在的数据集。此外，论文分析了各种分类器的范式，提出了多领域的识别任务、跨语言的识别任务、数据歧义问题等多方面的挑战，总结了未来 AIGC 识别的发展路径。旨在为相关科研人员提供清晰的介绍，为构建更加稳定高效的分类器提出建设性意见。

关键词：AIGC；机器生成内容检测；黑盒测试；白盒测试；深度学习

分类号：G251

生成式人工智能的快速发展，尤其是大语言模型的出现，为文本生成技术开辟了前所未有的新天地。OpenAI 公司推出的 ChatGPT 作为领域内的里程碑式作品，在许多专业工作流程中发挥了重要的作用，在故事生成[1-2]、广告标语生成[3]、新闻组成[4]、聊天对话生成[5]、代码生成[6]和放射学报告生成[7]等方面表现出了卓越的性能。同时，大语言模型凭借着其优秀的语义理解能力，在教育、医疗保健、商业、制造业等领域扮演了关键的角色，在提升工作效率、推动创新、促进跨文化交流等方面具有积极的影响。

由于大语言模型具有强大的文本生成能力，个人往往无法有效识别人工智能生成内容和人类生成内容，产生了许多伦理、社会和认识论方面的窘境。长期研究网络虚假有害信息的互联网公司 NewsGuard 联合 CEO 坦言：ChatGPT 将成为互联网上最强有力的散播虚假信息的工具[8]。“AI 教父”、图灵奖获得者辛顿更是发出了警告：生成式人工智能正在制造大量虚假的文本、图片和影像……若没有及时准备好相关法规和有效控制手段，人类在未来将对 AI 彻底失去控制[9]。以上言论并非危言耸听，多位科学家呼吁人工智能实验室停止训练更加强大的人工智能系统[10]。学术界也逐渐重视 AIGC（AI-Generated Content，人工智能生成内容）引发的问题，主要集中在两个方面。第一，AIGC 容易受到捏造信息、过时信息以及提示关键词的影响，这引发了错误信息[11-12]、学术不端[13]、钓鱼邮件[14]等问题，阻碍了 AIGC 在媒体和教育领域的发展。第二，人为恶意使用大语言模型，以极低的成本促进了虚假信息传播[15]、网络欺骗[16]和政治宣传[18]。经过行

---

*本文系国家自然科学基金青年项目“基于多源异构数据的科技关键节点及信息扩散机理研究”（项目编号：72304169）的研究成果之一。

作者简介：乔鸿，硕士生导师，副教授，博士研究生，E-mail：byant_sdnu@foxmail.com

为人有意的训练和提示后，生成式人工智能可以输出虚假有害的信息。AIGC的滥用，对信息生态环境造成了不良的影响[19]，针对这一问题，一种方式是通过人类的专业知识判断当前文本是否为AIGC。但是，人工识别效果不佳[20]，其准确率非常低，近乎等于随机分类的值。第二种方式是通过白盒检测的方法，白盒检测通常是由人工智能的开发人员创建分类器，它可以随时访问大语言生成模型，给AIGC打上标签，保证AIGC的可追溯性。第三种是通过黑盒检测的方法，利用人类生成文本和AIGC来训练分类器，与白盒检测相比有较高的稳定性，同时比人工识别的效率及准确率提高了20%-40%。因此，开发强大可靠的分类器来高效识别AIGC，降低生成式人工智能的滥用以及治理信息环境中的AI污染至关重要[22]。

以上问题最常见的解决方法是将人类文本与AIGC视为一个二分类问题，对此，自然语言处理领域做了一系列的努力。黑盒检测和白盒检测是其中主流的方法，但是，AIGC识别是涉及计算机科学、语言学、信息资源管理等多个领域的跨学科问题。现有的综述研究主要是针对当时新发布的几个对话生成模型的检测算法进行了梳理和分析[24]，其对检测方法的综述深度不足。同时也鲜有研究从自然语言处理视角出发，对AIGC识别的研究进展进行综述。

本文系统综述了自然语言处理在AIGC识别问题上的相关研究，旨在帮助相关学者应对未来出现的问题和挑战。本文试图解决以下研究问题：

1.介绍AIGC识别的任务，以及自然语言处理赋能下的AIGC识别研究热点聚焦在哪几个方面？

2.目前研究中所使用的数据集是否足够全面？存在的问题有哪些？

3.识别AIGC的分类器是如何发展的？各种类型分类器的局限性？

4.目前AIGC识别面临的挑战有哪些？如何应对这些挑战？

1.方法论/Methodology

1.1 研究方法

随着科学文章的不断增长，通过系统化的循证科学方法论对海量研究成果进行总结概括的综述类文献越来越多[25]。综述文章的迅速增长促进了综述方法的发展，这些方法的名称虽各不相同，但是都根据其需求解决某些特定的问题，见表1。例如，文献计量方法得出的结论可以帮助研究人员在海量文献中筛选出有价值的资源，了解某个学术领域的研究热点、学术影响力和学者的贡献，帮助学者评估学术成果的质量和影响[26]；元分析则是聚焦于不同的定量研究结果，经过权衡和比较，确定在同一主题下的多个研究中出现的范式、分歧或者关联[27]；扎根综述是对某个主题进行深入分析和解读，强调整个综述过程中的不断迭代，直至主题或理论的提出达到饱和；质性系统评价通过系统地收集、评估和合成质性研究的结果，提供对特定主题或现象的深入理解。系统综述法则是围绕更具体的研究主题，通过严格的文献筛选标准，确定研究主题的范围和纳入标准，更加适用于了解已有研究的现状和进展，发现研究中存在的挑战与问题。因此，本文选用此方法进行综述。

表1 **文献综述方法总结**

Table 1 Literature Review Methodology Summary

综述方法	目的	聚焦问题	特点

文献计量法[26]	来评估某个领域的学术研究水平、学术影响力以及学术趋势等信息	对后续研究的范围界定	具有丰富可视化功能的计量软件
元分析[27]	对研究结果的整合与比较	整合不同研究的数据	通过系统性的分析，重新评估定量结果
质性系统分析[28]	整合多项异质研究	批判性地评论和整合现有研究文献	整合同一主题的定量和定性研究
扎根综述法[29]	深入挖掘问题、现象或领域的内在本质和规律性	系统而严格的概括过程	通过上升归纳和理论饱和来生成新的理论或拓展现有理论
系统综述法[30]	对研究结果进行提炼和比较	揭示该领域的研究重点和挑战	根据具体研究问题制定相应的具体标准

系统综述方法收集和分析相关研究的论文数据[30]，进行识别和批判性评估。该方法在收集文章和研究证据时，使用系统、明确的方法，很大程度上减少了偏见因素的干扰[31]。其过程严谨、目的明确、可重复性强等特点，逐渐受到 AI 文献综述领域的重视。本文使用系统综述法的流程如下：①定义与检索。对研究问题进行定义与含义表述、确定文献来源。②筛选文献。根据入选标准进行文献筛选。③记录信息。从选定的文献中提取关键信息，如研究设计、数据集、评价指标等。④综合分析。根据纳入研究的特征、总结主要发现、讨论研究结果，并进行次分析或亚组分析，发现研究中存在的挑战与问题。

1.2 文献检索

首先，对人类生成内容（HGC，Human Generated Content）、人工智能生成内容（AIGC）、AIGC 识别进行阐释。将生成内容形式限定于文本类型，①人类生成内容：人类创作或产生的文本，这些内容可以是个人的作品，也可以是在社交媒体、论坛、博客等平台上发布的用户生成内容。HGC 通常反映了个体的观点、经验、创意和情感等，具有多样性和独特性。②AIGC：由计算机程序或人工智能技术生成的文本内容，通常是基于机器学习、自然语言处理、大语言模型等技术实现的，可以自动从大量数据中学习、推断和生成新的内容。③AIGC 识别：其本质为二值分类问题，目的是识别给出的文本内容是否由人工智能生成，其数学表达式为：

$$Y(x)=\begin{cases}1 & \text{if x is AIGC}\\0 & \text{if x is HGC}\end{cases} \tag{1}$$

其中，Y(x)是分类器，x 是需要识别的文本。

其次，根据上述阐释，立足于 AIGC 识别的研究主题进行概述，保证查全率的基础上拥有较高的查准率。制定检索词时，依据相关综述文章收集到合适的关键词，进行组合检索，并以相关综述文献的检索结果作为参考，最终选择了具备较高查全率和查准率的检索式：AIGC 识别 OR 机器生成文本识别 OR 大语言模型生成内容识别 OR 深度伪造内容检测。随后，在检索中，采用中文数据库中国知网（CNKI）、英文数据库 WoS、Google Scholar、ACL 等，检索数据库见表 2。检索时间为 2023

年 10 月 31 日。

表 2 检索数据库

Table 2 Search Database

数据库来源	数据类型	数量
CNKI	Topic	38
Google Scholar	Full Text	310
ArXiv	Full Text	539
WOS	Topic	500
IEEE Xplore	Full Text	836
Springer Link	Full Text	19
ACL	Full Text	N

注：在 ACL 数据库中不能使用所有检索词进行组合，有重复篇章存在，因此检索数量不准确。

1.3 文献筛选

为有效筛选文献，本文制定如下审查标准：①研究问题与本文研究目的有一致性，文献应该是关于 AIGC 识别的相关方法研究或者综述。②提出一种识别 AIGC 的方法、模型。③文章应该针对 AIGC 的相关研究指明前瞻性方向。只要文献满足以上条件中的一条，就将其收集到本文的研究数据中。

根据上述审查标准对文献进行筛选，①对检索到的文章进行学科限定，排除关联强度低的学科（如地理、历史、天文等），尽可能保留相似学科（计算机科学、信息科学、信息管理等），最终获得文献 2242 篇。②根据审查标准，对文献进行剔除。首先，通过阅读标题、关键词、摘要来排除不相关文献，最后根据全文内容确定是否选入文献集合，最终选择了 65 篇文献。③分析入选文献的参考文献，通过反向检索，对文献集合进行补充，获得最终文献集合 71 篇。

2.结果/Result

2.1 数据集综述

为高效识别 AIGC，学术界通过分类算法构建高效的分类器，提取数据中的有效信息，机器学习、深度学习成为解决该问题的关键技术。高质量数据集是训练良好的机器学习和深度学习模型的关键，这些模型需要大量的标记数据来学习并理解 AIGC 的特征和模式。丰富而有代表性的数据集能够提供多样化的样本，帮助模型更全面地学习并准确识别 AIGC。本节综述了识别 AIGC 任务的主流数据集，有助于研究人员了解目前可用的数据集，包括其特征、规模、质量以及针对 AIGC 识别的有效性。同时，针对现有数据集的不足，介绍了未来可能会使用到的数据集，提供制定更完善和全面的数据集标准，推动 AIGC 识别技术的发展，从而更好地应对海量的 AIGC 所带来的挑战。

当下研究中最流行的数据集如表 3 所示，作为 AIGC 识别任务的数据来源，这些数据集经常被用来测试相关算法的效率。未来研究中的潜在数据集如表 4 所示，这些数据集为跨场景下的 AIGC 识别任务提供了良好的数据来源。

表 3 主流数据集相关参数描述

Table 3 Mainstream Data Set Related Parameter Description

数据集	人类生成文本	AIGC	生成式人工智能类型	语言	文本领域
HC3[32]	58k	26k	ChatGPT	英文	开放领域、计算机、金融、医学、法律、心理学和许多其他领域

HC3-Chinese[32]	22k	17k	ChatGPT	中文	提供人类和人工智能的中文回复，为对话系统研究提供可比较的语料库
CHEAT[33]	15k	35k	ChatGPT	英文	由 ChatGPT 编写的大规模特征 数据集、包含 35,304 条合成摘要、 用于支持检测算法的开发。
CAPTION[34]	252.3k	932.3k	GPT-2	英文	该数据集是利用澳大利亚广播公司的头条新闻和模型生成的头条新闻创建的，并在头条新闻数据上对预先训练好的 GPT-2 模型进行了微调。
GROVER Dataset[35]	15k	10k	Grover-Mega	英文	该数据集包含 100 万对话，主要用于训练对话机器人以提高其对话能力和自然语言文本的生成能力。
TweepFake[36]	12k	12k	GPT-2, RNN, Markov, LSTM, CharRNN	英文	数据集包含真实新闻文章和假新闻文章，并使用许多不同的方法构建，包括基于语言建模的方法、基于机器学习的方法和基于深度学习的方法。
GPT-2 Output Datase[37]	250k	250k	GPT-2	英文	包含 250K 来自 WebText 测试集的文档，每个 GPT-2 模型（在 WebText 训练集中训练）有相应的 250K 随机样本（温度 1，没有截断）和使用 Top-K 40 截断生成的 250K 样本。
TuringBench[38]	10k	190k	GPT-1, GPT-2, GPT-3, GROVER, CTRL, XLM, XLNET, FAIR, TRANSFORMER XL, PPLM	英文	它包含多个任务的数据集， 包括图像分类、文本生成、语音识别等。每个任务都有一个相应的数据集，用于评估 AI 模型在该任务上的性能。

MGTBench[39]	2.5k	13k	ChatGPT, ChatGPT-turbo, Chat-GLM, Dolly, GPT4All, StableLM	英文	包含几种不同类型的数据集，涵盖了广泛的任务，如文本分类、语言翻译和文本生成。这使得评估大型语言模型在不同任务上的性能成为可能。
ArguGPT[40]	3.7k	3.7k	GPT2-xl, Text-babbage-001, Text-curie-001, Text-davinci-001, Text-davinci-002, Text-davinci-003, GPT-3.5-turbo	英文	该语料库在 4038 篇议论文中，有 7 个 GPT 模型根据了三个来源的论文提示生成：(1)课堂或家庭作业，(2)托福和(3) GRE 写作任务。
DeepfakeText-Detect-Dataset[41]	432.6k	-	GPT, LLaMA, GLM-130B, FLAN-T5, OPT, T0, BLOOM	英文	通过收集从各种人类著作中提取的文本和由不同的大型语言模型产生的深度伪造文本组成
M4[42]	123k	123k	ChatGPT, Textdavinci-003, LLaMa, FlanT5, Cohere, Dolly-v2, BLOOMz	英文 中文 俄文 印尼语、阿拉伯语	该数据集是一种用于机器生成文本检测的多生成器、多域、多语言的语料库。
GPABenchmark[43]	600k	600k	GPT-3.5	英文	这是一个基准数据集包含 60 万篇手写、gpt 编写、gpt 完成和 gpt 润色的计算机科学、物理、人文和社会科学研究论文摘要的样本。
Scientific-articles[44]	12k	12k	SCIgen, GPT-2, GPT-3, ChatGPT, Galactica	英文	此数据集包含来自 SCIgen、GPT-2、GPT-3、ChatGPT 和 Caladicga 的人类撰写和机器生成的科学论文
RCDataset[45]	20k	15.8k	ChatGPT	中文	人工智能生成的内容数据通过输入不同的提示到 ChatGPT 中获得，其中一部分通过从 THUCNews、WebQA 和 Moviedata 等数据集中选择作为人工生成的内容数据集。

HAC[46]	52k	48.9k	ChatGPT	中文	在百度问答、百度贴吧、新浪微博上获取人工问答，将答案作为人工回答数据，然后将问题输入 ChatGPT，让其模拟人工回答作为人工智能回答数据，然后筛选出人工智能生成内容。

潜在数据集中分为问答类、新闻类、学术论文类、社交媒体数据等四类。

问答类数据集包含了大量的问题和答案，可以用于训练模型，使其能够完成识别 HGC 和 AIGC 的任务。如表 4 所示，这些数据集中的问答涵盖了各种主题和领域，可以帮助模型学习各种知识和技能，从而更好地理解人类语言和机器生成内容的异同。

新闻数据集通常包含各种语言风格和表达方式，涵盖了丰富多样的内容，从政治、科技到娱乐等不同主题领域。这一多样性有助于训练模型更好地理解和生成各种类型的内容，并提高对真实信息的识别能力。相关学者偏爱将新闻标题输入到大语言模型中生成摘要，或者将摘要输入到大语言模型中让其生成标题，再将 AIGC 与 HGC 进行对比分析。

生成式人工智能广泛应用在学术写作中[55]，给大语言模型一个特定的学术主题，它可以高效的生成一篇论文或一段摘要，这些数据集为识别论文是否由人工智能生成提供了支持。

社交媒体用户发布了大量的文本内容，包括短文、评论和帖子等。这些数据可以用于训练大语言模型，以生成类似的文本。通常为生成式人工智能提供一个开始句，允许它们继续生成相关内容，或者根据社交媒体上的标题生成相应文本。

表 4 潜在数据集相关参数描述

Table 4 Description of Relevant Parameters of Potential Data Sets

数据集	大小	文本来源	语言	领域
ELI5[47]	556k	Reddit	英文	新闻文章、维基百科页面、问题回答
NarrativeQA[48]	1.4k	网页	英文	小说、电影剧本、问题回答
百度知道问答数据集[49]	100k	百度知道	中文	科技、生活、娱乐、健康等
知乎问答[50]	34.3k	知乎	中文	健康和医学、生命科学、地球科学等问答
PubMedQA[51]	211k	PubMed	英文	生物医学问答
Extreme Summarization (XSum)[52]	225k	BBC	英文	新闻、政治、体育、健康、家庭、教育、娱乐等多个领域
THUCNews[53]	740k	新浪新闻	中文	财经、彩票、房产、股票等多个领域
SogouCS[54]	2910k	搜狗新闻	中文	体育、金融、娱乐、汽车、技术等多个领域
PeerRead[56]	14.7K	NIPS,CoNLL,ACL,ICLR,arXiv	英文	论文草稿和一些顶级会议接受或拒绝的科学论文
ArXiv[57]	2300k	ArXiv	英文	物理学、数学、计算机科学与生物学等学科论文
Chinaxiv[58]	38k	Chinaxiv	中文	涵盖物理、天文、生物、图书情报等学科论文
WebText[59]	45m	网页	英文	抓取网页数据，并把其他数据集的通用数据源删

				除
Avax Tweets Dataset[60]	137 m	推特	英文	新冠肺炎帖子
IMDB Dataset	50k	电影数据	英文	电影评论
Yelp[61]	700k	Yelp	英文	企业评论

2.2 分类器综述

本部分介绍自然语言处理技术如何赋能 AIGC 识别。主要分为两类方法，一类是白盒检测中的水印技术；另一类是黑盒检测方法中的零样本分类器、微调 LMs（Language Models）分类器、LLMs（Large Language Models）作为分类器的方法，各种方法的原理流程图如图 1 所示。

2.2.1 白盒检测

在白盒检测中，分类器会将隐藏的水印添加到大语言模型生成的内容中，为后续的追踪工作提供监察，防止其进行危害社会的活动。白盒方法最早出现在计算机视觉领域的生成模型的开发，现已加入 AIGC 检测的行列，其优点在于能够保护 AIGC 的版权、确认生成内容是否被篡改或修改并且能够追踪到内容的来源，抵抗攻击性较强，同时优秀的水印技术可以在不影响原始内容质量的情况下，嵌入隐蔽的水印信息。

（1）统计分析方法

这类方法通过对水印文本和非水印文本之间的输出标记进行统计分布分析，寻找其输出标记与逻辑统计上的差异，从而推断文本中可能存在的水印。依靠统计学原理和数学模型，解释差异是如何反映水印存在的可能性，使检测者能够理解检测结果的依据。John Kirchenbauer 等[62]提出了一种针对大语言模型的专有水印框架，该框架可以嵌入水印，同时对文本质量的影响可以忽略不计，并且可以使用高效的开源算法来检测，而无需访问语言 API 模型或参数。其原理在于，生成单词之前选择一组随机的“绿色”标记，这意味着其它的标记为“红色”，然后在采样期间温和地促进绿色标记的使用。该研究还加入了基于 P 值的可解释性模块，在最后分类阶段可以将绿色标记和红色标记进行统计，计算 P 值然后确定生成的水印。最近的一项工作中[63]，引入了“WinMax”的窗口测试，探讨了水印文本在人类重写、使用非水印 AIGC 的转述或混合到更长的手写文档中后的稳健性。该研究主张将水印可靠性作为文本长度的函数，同时发现即使是人类作者也不能可靠地去除水印，说明了水印方法的可行性。

（2）密钥的水印技术

Zhao 等人[64]提出了抗蒸馏水印(distill- resistant Watermarking，DRW)方法，将水印嵌入到模型产生的预测概率向量中。该嵌入与密钥相对应，可以保护模型并且通过探测可疑模型来检测密钥信息。后者有助于确定可疑的模型是否从受保护的模型中提取出来。DRW 方法的使用有效保护了 NLP 模型免受未经授权的蒸馏，同时保持了它们的准确性和完整性。因此，DRW 提供了一个强有力的机制来保护包含在精心训练的 NLP 模型中的知识产权，防止潜在的盗窃和滥用。Liu 等人[65]提出了第一个私有水印算法，创新性地使用两个不同的神经网络进行水印生成和检测，而不是在两个阶段都使用相同的密钥。同时，令牌嵌入参数在生成和检测网络之间共享，有效提高了精准度且对生成和检测过程的速度影响都很小。

2.2.2 黑盒检测

黑盒检测方法仅限于对 LLMs 的 API 级访问。它通过从人类和机器来源收集

文本样本以训练分类模型，该模型可用于区分 ChatGPT 生成的文本和人类生成的文本。比如，Dugan 等人[66]通过人工方式构建数据集，用以评估自然语言生成系统的质量，衡量人们对生成文本的感知。此外，Guo 等人[32]通过整合维基百科等现有的问答数据集，通过预训练模型微调，研究了人类文本与 AI 文本的各自特点以及相似度。黑盒检测一般由外部实体构造，不需要了解其具体的工作机制。这使得检测方法更具通用性和适用性，即使在无法获取模型细节的情况下，也能进行有效的检测。

（1）Zero-shot Methods

该方法与水印方法不同，它可以通过分析文本的特征和统计数据进行分类，Simon Corston-Oliver 等[67]是最早开创零样本检测研究的学者，他们提出了一种机器学习方法来评估文本是否是机器翻译系统输出的，基于语言特征的分类器来识别人类翻译和机器翻译的文本，这些语言特征例如分支属性、虚词密度和组成部分长度是区分这两类文本的关键因素。还有学者使用频率统计的方法完成类似的任务，Leonid A 等[68]学者通过词频统计机制，来识别文本是否自动生成。Yuki Arase 等[69]专注于在现有自动检测统计机器翻译（SMT）结果中观察到的短语沙拉现象，提出了一组计算简便的特征来有效地从大规模 Web 挖掘文本中检测机器翻译的句子。Perplexity 的方法是基于传统的 n-gram 语言模型[70]，通过困惑度来评估语言模型识别文本的熟练程度，使用 SRILM 工具包来计算困惑度的值[71]。最近广泛流传的 GPTzero 基于对文本的困惑度和突发性度量进行了深入研究，在识别 AIGC 方面效果较好[72]。熵也是一种早期的 Zero-shot Methods，通过 Kullback-Leibler (KL)散度对 n-gram 进行评分，考虑了单词之间的距离信息，有助于虚假内容的识别[73]。同时，基于 Log Rank 的方法也不断浮出水面，利用大语言模型来分析文本中的单词排名，通过比较文本中词汇使用分布情况和大语言模型中词汇使用分布情况，判断文本由大语言模型生成的概率。GLTR[74]基于以上原理设计而成，将对比过程进行了可视化，根据单词的不同频率进行不同颜色的标记，通过鲜明的颜色突出大语言模型在生成文本时倾向输出单词的概率。DetectLLM[75]作为 Zero-shot Methods 中最先进的分类器，在 GLTR 的基础上引入了 LRR（Log-Likelihood Log-Rank Ratio），使得该分类器显著提高了效能。但是 DetectLLM 方法的鲁棒性较差，如果对经过审查的文本进行扰动的过程中无法保持语义相似度，分类器的性能会出现大幅度下降。同时，因其需要对多个扰动进行评分，评估过程花费时间成本过高，这也是制约 DetectLLM 发展的一个重要原因。为了减少时间成本，有学者使用贝叶斯代理模型对评分过程进行了改进[76]，通过选取少量典型样本进行评分，然后将分数插入到其它样本中提高查询效率，在保持性能的同时降低了一半的时间成本。也有学者为减少 DetectGPT 密集的计算成本带来的损失，引入了条件概率曲率，提出了 Fast-DetectGPT[77]，其精度提高了大约 75%，检测效率提高了 340 倍。

（2）Fine-tuning LMs Methods

通过微调基于 transformer 的生成式人工智能，来区分由人工智能生成的文本和非人工智能生成的文本。经过训练后的模型在自然语言理解方面有了非常大的提升[78]，而模型的自然语言理解能力对文本分类任务极其重要。一些优秀的预训练模型，例如 BERT[79]、Roberta[80]和 XLNet[81]，在 GLUE 基准中应用于文本分类任务时，在传统的统计机器学习和深度学习方面表现出了优于同类模型的性能[82]。同时，大量研究已经证明 Fine-tuning LMs Methods 在 AIGC 识别方面具有强大的能力[41]，特别是 Roberta[80]，是 AIGC 识别任务中最优秀的分类器之一。微调

的 Roberta 为 AIGC 识别任务提供了鲁棒性较高的基准，Fagni T 等[36]使用三种不同的方法对文本内容进行编码，最终发现微调的 Roberta 是分类效果最好的方法，OpenAI 公司公布的分类器也采用了微调 Roberta 的方法[59]。在 AIGC 识别任务方面，这些基于 BERT 进行微调的模型，与 Zero-shot 和白盒方法相比有着惊人的正确率，其平均正确率达到了 95%以上。同时，在特定领域内具备抵抗各种攻击技术的能力，不易受到攻击的影响和破坏。但是，当数据变为跨领域数据集或未知数据时，其性能开始大幅度下降，在识别不同语言模型生成的数据时效果较差[85]。鲁棒性较差是 Fine-tuning LMs Methods 的通病[41]，因为这些方法都过度拟合于所训练的数据集，导致面对跨领域数据集和未知数据时，性能大幅度衰退。

（3）LLMs 分类器

为了对抗大语言模型生成的虚假信息，Rowan Z 等[35]最早提出了 LLMs 作为分类器的构想，他们设计了一个文本生成模型 Grover 用来生成文本信息，因为 Grover 固有的可控性质，该模型生成的信息具有显著的欺诈性。他们针对 Grover 生成的信息使用多种分类器（BERT[79]等）进行识别，发现最好的分类器的准确率为 73%。出人意料的是，抵抗 Grover 生成文本的最好分类器是 Grover 本身，具有 92%的准确率，这也说明了开发强大生成器的重要性。但是 LLMs 分类器的可靠性是一直被怀疑的，有学者对 ChatGPT 和 GPT4.0 等主流的生成式人工智能作为分类器的效果进行了研究，发现这些 LLMs 识别 AIGC 的可靠性非常差[86]。但是，ChatGPT 与 GPT4.0 有着完全相反的表现。ChatGPT 作为分类器识别 AIGC 的准确率不到 50%，这也说明了 ChatGPT 无法在大量文本中识别出 AIGC；有趣的是，ChatGPT 在识别 HGC 方面表现更好，并且倾向于将 AIGC 分类为人类生成文本。GPT4.0 几乎将所有文本都归类于 AIGC，这说明 GPT4.0 难以识别 HGC。Liu 等[40]测试了 GPT3.5-Turb 分类器的效果，在 zero/few-shot 背景下的分类准确率均低于 50%。这些研究均表明 LLMs 并不是一种可靠的分类器，其分类准确度远低于水印方法和其它的黑盒方法。为了有效利用 LLMs 的强大能力，Yu 等[87]引入了 GPT-Pat，缓解了 LLMs Methods 普遍性缺乏的劣势和 LLMs 的不可靠性。GPT-Pat 对识别文本的源问题进行溯源，然后根据推断出的源问题重新生成文本，最终计算识别文本与生成文本的相似性，这为 LLMs 识别 AIGC 提供了新思路[87]。该方法不仅拥有良好的分类效率，其鲁棒性也非常优秀，对于改写与润色的人工智能生成文本也有一定的抵抗能力，其性能衰退率仅为 Rooberta[80]的一半。但是，该方法在训练和识别期间查询 ChatGPT，这导致用户需要花费大量的时间成本，用户体验较差。

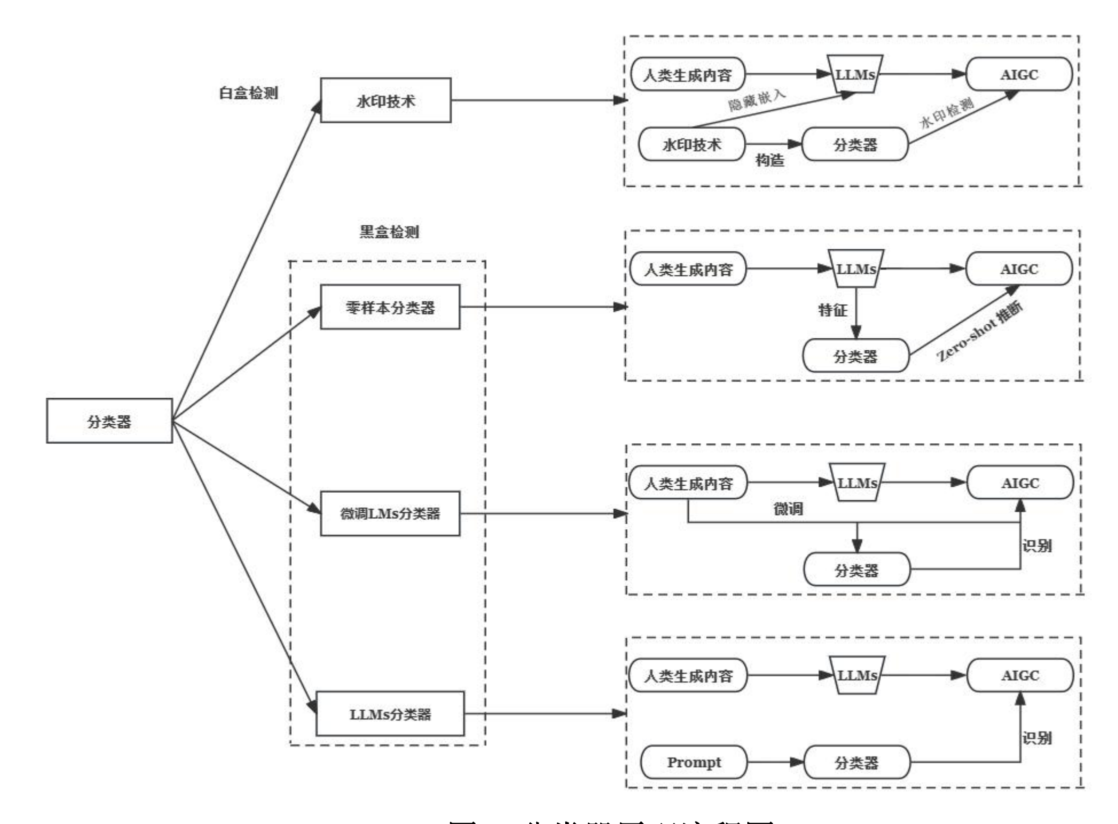


图 1 分类器原理流程图

Fig. 1 Classifier Principle Flow Chart

2.3 评价指标

评价指标能够量化模型在 AIGC 识别任务中的表现，是任何 NLP 任务的必要组成部分。我们列举了 AIGC 识别任务中常用的指标，为后续评价分类器的效果提供了度量。

AIGC 识别任务所有可能的混淆矩阵类型只有四种：

（1）阳性（True Positive，TP），如果原始内容是 AI 生成的，并且分类器正确地将其分成 AI 生成文本，则分类器的响应被归类为阳性。

（2）阴性（True Negative，TN），如果原始内容是人类生成的，并且分类器正确地将其分成人类生成文本，则分类器的响应被归类为阴性。

（3）假阳性（False Positive，FP），如果原始内容是人类生成的，并且分类器错误地将其分成 AI 生成文本，则分类器的响应被归类为假阳性。

（4）假阴性（False Negative ，FN），如果原始内容是 AI 生成的，并且分类器错误地将其分成人类生成文本，则分类器的响应被归类为假阴性。

下列分类性能指标均可以用 TP、TN、FP、FN 来表示，包括准确率（Accuracy）、精准率（Precision）、召回率（Recall）、F1 值。

准确率（Accuracy）是一个通用的度量值，是评估模型在分类问题中整体预测正确的能力指标。准确率的计算公式如下：

$$\mathrm{Accuracy} = \frac{\text{正确分类数量}}{\text{所有文本数量}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (2)$$

精准率（Precision）:评估模型在预测正类别样本中准确性的指标。它衡量了模型在预测为 AIGC 的样本中，实际上有多少是真正的 AIGC 类别样本。公式如下：

$$\text{Precision} = \frac{\text{正确分类的 AIGC 数量}}{\text{所有检测到的 AIGC 数量}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

召回率（Recall）：评估模型在所有正类别样本中成功预测为正类别的能力指标。它衡量了模型识别出的正类别样本占实际正类别样本的比例。平均召回率（AvgRec）是多类别分类问题中的一种综合指标，用于评估模型对不同类别的召回率表现。在 AIGC 识别的任务中，主要分为 HumanRec 和 AIGCRec，分别表示分类器准确分类为人类生成和人工智能生成的比例[88]。公式如下：

$$\text{HumanRecll} = \frac{\text{正确分类的 HGC 的数量}}{\text{所有 HGC 的数量}} \quad (4)$$

$$\text{AIGCRecll} = \frac{\text{正确分类的 AIGC 数量}}{\text{所有 AIGC 的数量}} \quad (5)$$

$$\text{AvgRecll} = \frac{\text{HumanRecll} + \text{AIGCRecll}}{2} \quad (6)$$

F_1值：一个综合评估模型性能的指标，是精确度和召回率的调和平均数，同时考虑了模型的精准率和召回率。公式如下：

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

3.讨论/Discussion

3.1 针对数据集构建的讨论

在 AIGC 识别问题上，专门为解决此问题所设置的数据集面临着诸多挑战，一个明显的趋势是利用解决其他任务的数据集来解决 AIGC 识别问题，有的在此基础上添加 AIGC 作为检测器的训练数据。形成这种趋势的原因在于本领域内没有专门解决 AIGC 识别任务的数据集，说明该任务的基准数据集不够全面、专业。未来数据集的构建应该具有以下标准：

全面性：一个合格的数据集应包含不同领域的、不同任务的、多语言的数据内容，这样才能促进识别 AIGC 效率更高的分类器的产生。配置不同领域的数据内容，对提高分类器的鲁棒性、可信度具有重要意义。日常识别任务中，分类器应该可以识别学术论文、新闻标题、微博问答等多个场景下的文本内容。同时还应配置不同语言的数据集，为探索跨语言分类器的发展提供基础。在不同的语种下分类器可能会产生不同的效果，这会制约分类器的发展，全面的数据集为分类器的高速发展提供了方向。另一方面，分类器会遭遇多种机制生成文本的攻击，因此数据集中还应该包括各种攻击文本（经改写、润色、替换同义词处理过的文本），这有助于提升分类器的有效性。

时效性：从上述数据集的综述中可以看出，一些非常久远的其它任务中的数据集作为 AIGC 识别的数据来源，这导致训练分类器的数据内容可能是过时的，这也意味着经过训练的分类器可能跟当前社会中的内容脱轨，导致分类器在现实中的效果并不好。因此，我们需要建立更新的数据集，保证训练分类器的数据能够与时俱进。

多样性：大语言模型的快速发展，其产品类型也多种多样，国内外有文心一言[90]、盘古[91]、LaMDA[92]、PaLM[93]、Jurassic[94]等多种大语言模型，其生成内容与 ChatGPT 相比各有优劣。但在学术界对大语言模型生成文本检测中，较多数据

集使用的人工智能生成文本都是由 ChatGPT 生成的，不利于分类器的发展。同时，不能忽略其它大语言模型所带来的挑战与风险，我们应该从不同大语言模型中收集数据，构建具有多样性的数据集，使得分类器可以识别多数大语言模型生成的文本，并且能够抵抗不同语言模型生成文本的攻击。

3.2 针对分类器设计的讨论

没有哪一种分类器是万能的，我们所讨论的局限性问题是整体分类器所面临的困境。

多领域的识别任务：跨领域使用某种模型是整个 NLP 领域的重大难题，不同领域的语言和术语使用方式可能差异很大，模型需要适应并理解多种不同领域的语言表达。例如，医学领域的术语和金融领域的术语可能完全不同，导致 GLTR[74]、DetectLLM[75]、微调 Roberta[80]等方法在面对跨领域识别任务时，性能出现了显著下降，这也凸显了开发跨领域分类器的必要性。实际上，跨领域的分类器需要大量且多样化的数据来学习各个领域的语言特征，然而，获取特定领域的大规模数据并非易事，有时候某些领域数据可能非常有限。因此，可以通过迁移学习、领域自适应、多任务学习以及对模型架构的改进，来促进分类器在多领域进行识别任务的效率提升。

跨语言的识别任务：不同语言之间存在巨大的差异，包括语法、词汇、句法结构等。跨语言模型需要能够理解和处理这些差异，并且具备足够的灵活性。Yuxia[42]与 Chaka[95]研究发现，不同语言可能存在一定的迁移能力，但是在多种语言的分类器中发现，在识别非训练数据语种的内容时，准确率出现了下降。最新的研究中也发现了这个问题[96]，面对非英语母语者撰写的文本时，最先进的分类器的性能出现了明显的下降。通过使用有效的提示策略可以缓解这种问题，但它也会增加生成文本逃过分类器的概率。这也表明分类器可能会在检测过程中出现歧视问题，当数据集中有非标准语言的文本时，分类器会惩罚此类文本，导致分类器性能下降[96]。

数据歧义问题：大语言模型在生成文本时，可能缺乏足够的上下文，使得模型产生不完整或不准确的信息。这可能导致歧义的产生，因为模型无法正确理解或完整把握输入的含义。如果无法分辨此类信息是人类生成的还是机器生成的，并且将 AIGC 当作模型的预训练数据，就会导致恶性循环，致使分类器识别效率大幅度下降，破坏了分类器最初的任务前提。

3.3 针对评价指标的讨论

评价指标是评估模型性能不可或缺的部分，我们讨论了一般性评价指标和新型评价指标的优缺点，并创建了 AIGC 分类器评估框架，进一步启发相关分类器的研究，为后续不同背景下的研究提供合适的指标。

3.3.1 一般性评价指标

准确率（Accuracy）适用于数据类别分布均衡的数据集，能够直观地观察到模型整体预测正确的比例。但是，对于数据类型分布不均衡的数据集，其效果较差。平衡的准确率和不平衡的准确率在不同背景下得到了应用[35]，用来评价不同分类器在不同背景下的能力强弱。在 AIGC 识别任务中，数据集中 AGIC 样本数量应高于人类生成样本的数量，同时分类器识别 AIGC 的概率必须要高于识别 HGC 的概率。

精确率（Precision）在需要尽量避免误报的应用背景中（如癌症检测）是一个重要指标。当一个样本不属于 AIGC，而被分类为 AIGC，这个错误的结果会降低用户对模型的信任，对识别任务造成较大负面影响。当模型过度关注精确率

时，可能会牺牲召回率，导致漏报。

召回率（Recall）评估模型对真实 AIGC 样本的识别能力，在类别不平衡的数据集中，Recall 可以更好地反映模型在少数类别上的性能，避免过度关注常见类别而忽略了重要的少数类别。在数据不平衡的情况下，高 Recall 可能是因为模型倾向于预测更多样本为正例，而这些样本可能是错误分类的。因此，需要引入 HumanRecll、AIGCRecll、AvgRecll 等指标，综合评估模型的召回率。

3.3.2 新型评价指标

阴性预测值（Negative Predictive Value，NPV）是统计和诊断测试中阴性结果中实际为阴性结果的比例。在这种情况下，它表示在模型预测为 HGC 的情况下，真正的人类生成样本有多少被正确预测出来。当数据集中的类别不平衡时，NPV 可以提供更全面的模型性能评估，因为它们关注了特定预测类别的准确性。其公式为：

$$\mathrm{NPV}=\frac{\text{正确分类的 HGC 的数量}}{\text{所有分类为 HGC 的数量}}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \tag{8}$$

真阴性率(True Negative Rate ,TNR):有时也被称为特异性(Specificity)，是用于评估模型在所有实际的人类生成样本中,成功识别出的人类生成样本所占的比例[88]。公式如下：

$$\mathrm{TNR}=\frac{\text{正确分类的 HGC 的数量}}{\text{所有 HGC 的数量}}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{9}$$

假阳性率（False Positive Rate ，FPR）：用于衡量人类生成的样本被错误地分类为人工智能生成的样本比例。公式如下：

$$\mathrm{FPR}=\frac{\text{错误分类成 AIGC 的数量}}{\text{所有 HGC 的数量}}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} \tag{10}$$

假阴性率（False Negative Rate ，FNR）：用于衡量实际是人工智能生成的样本但被错误地分类为人类生成的样本比例。公式如下：

$$\mathrm{FNR}=\frac{\text{错误分类的 HGC 的数量}}{\text{所有 AIGC 数量}}=\frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} \tag{11}$$

AUROC（Area Under the Receiver Operating Characteristic Curve）：由 Receiver Operating Characteristic 曲线推导而来，用于衡量模型在不同阈值下的性能表现。公式如下：

$$\mathrm{AUROC}=\int_0^1 \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\, d \quad \frac{FP}{FP \qquad +TN} \tag{12}$$

3.3.3 针对 AIGC 分类器评估框架的讨论

为准确评估分类器在识别 AIGC 方面的能力，完善分类器的评估指标，我们构建了 AIGC 分类器评估框架。从分类器的应用需求出发，分析了分类器与各类 AIGC 之间的内在关联，构建了系统性评估 AIGC 分类器的框架思路。目前针对分类器识别 AIGC 的能力评估只有一些孤立的评价指标，学界和业界还没有形成成熟的分类器评估框架。与此相反，在教育学中，已经有成熟的评估框架来指导教学任务，布鲁姆教学分类法（Bloom's Taxonomy）是其中较为广泛使用的框架之

一[97]。该分类法包括六个认知学习层次，按照认知复杂度逐渐增加的顺序排列：记忆（Remember）、理解（Understanding）、应用（Apply）、分析（Analyze）、评价（Evaluate）、创造（Create），如图 2 所示。

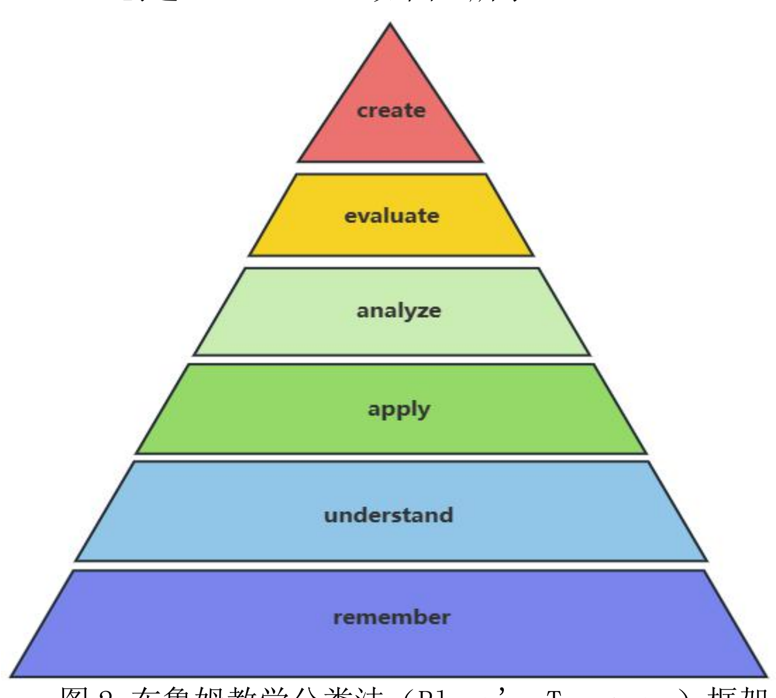


图 2 布鲁姆教学分类法（Bloom’s Taxonomy）框架

Fig. 2 Bloom 's Taxonomy Framework

受到布鲁姆教学分类法框架和 AIGC 分类器现实应用情况的启发，我们提出 AIGC 分类器评估框架来综合评估分类器的能力水平。该框架将 AIGC 分类器划分为四个能力层级，如图 3 所示，分别是学习、理解、识别、伦理四个层次。学习层次：评估分类器能否准确理解和记忆输入的内容。理解层次：评估分类器对 HGC 和 AIGC 的特征理解，这包括对不同来源的生成文本中上下文、语境和语义的理解。识别：评估分类器能否正确识别 HGC 与 AIGC，同时评估在面对抵抗攻击时，分类器的稳定性是否受影响。伦理：评估分类器在面临伦理和道德问题时，

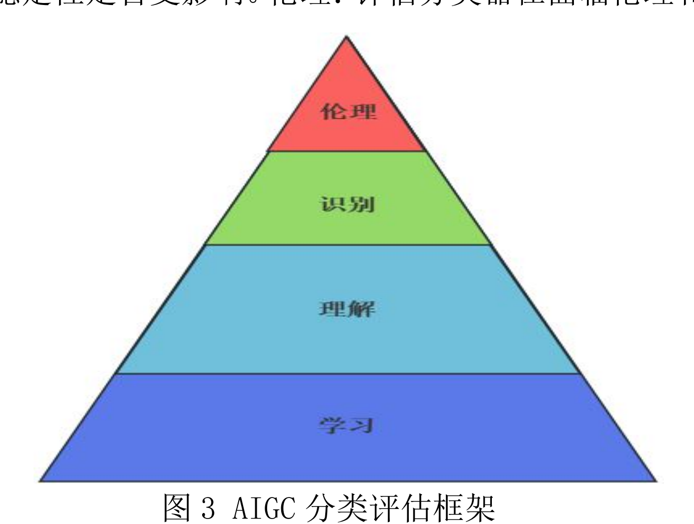


图 3 AIGC 分类评估框架

Fig. 3 AIGC Classification Evaluation Framework

是否会因为语言、种族偏见而引起分类器的性能下降。在评估过程中，也要考虑到数据的质量、模型的可解释性以及对潜在偏见和不公平性的审查，这些也是评估 AIGC 分类器非常重要的方面。

3.4 未来热点方向

AIGC 识别领域的相关研究已经取得了显著进展，但仍有一些问题存在。本小节探讨了未来研究的潜在方向，旨在推动构建更高效和实用的分类器。

3.4.1 重视类别不平衡的影响

目前为止，探究类别不平衡的研究相对较少。在现实生活中，AIGC 可能是少数类，分类器会受到严重的类别不平衡影响，导致性能不稳定或骤降[98]。使用单分类方法可能是解决 AIGC 识别问题的好方法[99]。针对 AIGC 识别任务，传统的监督学习面临着数据不平衡的挑战，可能将其误分类为正常类别。单分类方法专注于仅有一个类别的数据，并将其视为“正常”类别，而无需明确地定义其他类别。通过建模正常类别的分布和特征，单分类器能够更好地识别不同于训练集的异常样本。对于 AIGC 的识别问题，单分类方法可以更好地适应于少数类别，从而提高模型的泛化能力和鲁棒性[99]。

3.4.2 提高零样本检测方法的性能

零样本检测方法不仅稳定性高[76]，还可提供具有可解释性的结果[100]。AIGC 和人类生成文本之间有着明显的差异，GPT-4 生成内容中的高频搭配比 HGC 更多，同时它也惯用总结语句，有较强逻辑性[101]。这促进了 AIGC 识别的研究，我们应该深入探究 AIGC 与 HGC 之间的细微差别，从低维特征到高维特征聚焦各自特点。这样可以为分类器的构造提供准确的度量标准，为模型的决策过程提供可解释性。

3.4.3 构建能够抵抗对抗性攻击的分类器

对抗性攻击是阻碍当前分类器推广使用的主要阻力[39]，也是导致当前 AIGC 分类器持续不可靠性的一个重要因素。它通过对文本截断、打乱、单词交换和拼写错误对文本特征进行对抗攻击，这对微调分类器[41]、水印技术[62]、DetectGPT[98]构成了有效的攻击，检测器性能分别降低了 18%、10%和 25%以上。已有研究解决了特定攻击的鲁棒性问题，但是却忽略了其它类型的攻击所带来的潜在影响[102]。因此，必须开发和验证各种类型的攻击模式，检测现有分类器存在的漏洞。我们建议通过上文提到的 AIGC 分类评估框架，对现有分类器进行评估，最终构建出能够抵抗多种对抗性攻击的分类器。

3.4.4 检测方法要具有公平性和可解释性

使用黑盒检测方法进行 AIGC 识别任务，防止大语言模型的滥用，检测结果可能会给个人带来负面影响（学位论文造假、学术不端等）。AIGC 检测系统必须要保持适当的公平、透明和可解释性，要重视有关 AIGC 检测带来的潜在危害的技术或社会认知影响研究，这对确保检测系统的伦理性很重要。

可信任的人工智能政策要求决策系统提供人类可以理解的解释，并且反映在众多新兴技术监管指南和标准中[103]。有学者利用随机森林模型和 XGBoost 来检测 GPT-2 生成的虚假评论，并在分类器中加入了可解释性模块(Shapley Additive commentary, SHAP）[105]。未来应聚焦于识别效果好并且具有可解释性的识别方法研究。

3.4.5 防止检测过程中的歧视问题

在实际检测过程中，某些群体（如非母语人士）生成的文本更有可能被机器识别算法标记为 AIGC，这可能是由于他们的写作特征或使用翻译工具导致的[96]。为避免分类器产生歧视问题，首先要确保数据集中各个类别（不同语言或文化）的样本分布均衡，其次要使用公平性指标来评估分类器的性能，包括对不同语言和文化的分类的准确率。同时，还要监测分类器在不同类别数据上的表现差异，以及是否存在歧视的情况。未来如何在防止歧视的同时提高识别效率，是非常重

要并且具有挑战性的问题。

3.4.6 使用多智能系统辅助 AIGC 识别任务

有研究证明了多智能系统在提高 AIGC 识别性能方面的有效性，通过促进智能主体之间的协作和知识交换，利用多智能系统可以增强大语言模型的性能[106]。这种模式反映了人类的集体决策，Uchendu A 也证明了人类的集体决策在改进 AIGC 识别任务方面的能力[107]。因此，通过多智能系统辅助 AIGC 识别非常具有前景，多智能辅助系统可以利用智能体之间的集体商议来控制大语言模型生成内容的质量，也可以通过一种共识驱动的方式评估多个分类器的输出结果，从而产生更令人信服的结果。

4.结论与未来工作/Conclusion and Future Work

随着大语言模型技术的飞速发展，AIGC 在日常生活中无处不在。AIGC 的滥用导致了虚假评论、学术不端等问题。在此背景下，区分文本是人类生成还是人工智能生成具有重要意义。为了更好地防止 AIGC 滥用，减少虚假消息、学术不端、欺骗评论等问题的产生，本研究介绍了 AIGC 分类器的任务，指出了分类器的发展是时代的必然要求。为了设计出高效的分类器，对当前主流的数据集进行了介绍，指出了当前数据集存在的局限性，并且探索了未来可能作为 AIGC 识别任务的潜在数据集。此外，还阐明了当前识别 AIGC 的两种主流方式，白盒检测和黑盒检测，并且对分类器局限性进行了探讨，包括多领域的识别任务、跨语言的识别任务、数据歧义问题等。本文工作为为研究人员提供了清晰和全面的介绍，也希望能够为未来自然语言处理技术赋能 AIGC 识别任务激发了新的思路，促进更加高效的分类器的发展。

当然，AIGC 识别领域存在着大量的开放问题，比如，要建立全面的、多领域、多语言、先进的数据集。在此基础上，还要探索更先进的自然语言处理模型，通过对抗性攻击不断提高分类器的效率和鲁棒性;同时也要重视识别过程中的公平性和可解释性问题，减少识别过程中的歧视问题

参考文献：
[1] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
[2] Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation[EB/OL].(2018-05-13). https://doi.org/10.48550/arXiv.1805.04833
[3] Murakami S, Hoshino S, Zhang P. Natural Language Generation for Advertising: A Survey[EB/OL].(2023-06-22).https://doi.org/10.48550/arXiv.2306.12719
[4] Yanagi Y, Orihara R, Sei Y, et al. Fake news detection with generated comments for news articles[C]//2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES). IEEE, 2020: 85-90.
[5] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation[EB/OL]. (2020-05-02). https://doi.org/10.48550/arXiv.1911.00536
[6] Solaiman I, Brundage M, Clark J, et al. Release strategies and the social impacts of language models[EB/OL].(2019-11-13).https://doi.org/10.48550/arXiv.1908.09203
[7] Liu G, Hsu T M H, McDermott M, et al. Clinically accurate chest x-ray report generation[C]//Machine Learning for Healthcare Conference. PMLR, 2019: 249-269.
[8] Hsu T, Thompson S A. Disinformation Researchers Raise Alarms About AI Chatbots[J]. International New York Times, 2023: NA-NA.
[9] The New York Times.See Cade Metz,The Godfather of AI'Quits Google and Warns of Danger Ahead[EB/OL].(2023-05-12).https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html
[10] Reuters.Elon Musk and others urge AI pause, citing 'risks to society' [EB/OL].(2023-04-05).,https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/
[11] Shu K, Wang S, Lee D, et al. Disinformation, misinformation, and fake news in social media[M]. Cham: Springer International Publishing, 2020.
[12] Stiff H, Johansson F. Detecting computer-generated disinformation[J]. International Journal of Data Science and Analytics, 2022, 13(4): 363-383.
[13] Lee J, Le T, Chen J, et al. Do language models plagiarize?[C]//Proceedings of the ACM Web Conference 2023. 2023: 3637-3647.
[14] Baki S, Verma R, Mukherjee A, et al. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 469-482.
[15] Uchendu A, Le T, Shu K, et al. Authorship attribution for neural text generation[C]//Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020: 8384-8395.

[16] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models[EB/OL].(2021-12-08).https://doi.org/10.48550/arXiv.2112.04359
[17] Ayoobi N, Shahriar S, Mukherjee A. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention[C]//Proceedings of the 34th ACM Conference on Hypertext and Social Media. 2023: 1-10.
[18] Varol O, Ferrara E, Davis C, et al. Online human-bot interactions: Detection, estimation, and characterization[C]//Proceedings of the international AAAI conference on web and social media. 2017, 11(1): 280-289.
[19] Kreps S, McCain R M, Brundage M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation[J]. Journal of experimental political science, 2022, 9(1): 104-117.
[20] Dou Y, Forbes M, Koncel-Kedziorski R, et al. Is GPT-3 Text Indistinguishable from Human Text?[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.2022: 7250–7274.
[21] Clark E, August T, Serrano S, et al. All that's' human'is not gold: Evaluating human evaluation of generated text[C]// In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processin. 2021:7282–7296.
[22] Shevlane T, Farquhar S, Garfinkel B, et al. Model evaluation for extreme risks[EB/OL].(2023-09-22). https://doi.org/10.48550/arXiv.2305.15324
[23] Porsdam Mann S, Earp B D, Nyholm S, et al. Generative AI entails a credit–blame asymmetry[J]. Nature Machine Intelligence, 2023: 1-4.
[24] Crothers E, Japkowicz N, Viktor H L. Machine-generated text: A comprehensive survey of threat models and detection methods[J]. IEEE Access, 2023.
[25] Snyder H. Literature review as a research methodology: An overview and guidelines[J]. Journal of business research, 2019, 104: 333-339.
[26] Braun V, Clarke V. Using thematic analysis in psychology[J]. Qualitative research in psychology, 2006, 3(2): 77-101.
[27] Davis J, Mengersen K, Bennett S, et al. Viewing systematic reviews and meta-analysis in social research through different lenses[J]. SpringerPlus, 2014, 3(1): 1-9.
[28] 孙玉伟,成颖,谢娟.科研人员数据复用行为研究:系统综述与元综合[J].中国图书馆学报,2019,45(03):110-130.DOI:10.13530/j.cnki.jlis.190026.
[29] Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots[J]. International Journal of Human-Computer Studies, 2021, 151: 102630.
[30] Liberati A, Altman D G, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate

health care interventions: explanation and elaboration[J]. Annals of internal medicine, 2009, 151(4): W-65-W-94.

[31] Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement[J]. Annals of internal medicine, 2009, 151(4): 264-269.

[32] Guo B, Zhang X, Wang Z, et al. How close is chatgpt to human experts? comparison corpus, evaluation, and detection[EB/OL].(2023-01-18). https://doi.org/10.48550/arXiv.2301.07597

[33] Yu P, Chen J, Feng X, et al. CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts[EB/OL].(2023-04-24).https://doi.org/10.48550/arXiv.2304.12008

[34] Maronikolakis A, Schutze H, Stevenson M. Identifying automatically generated headlines using transformers[C]//In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda.2021:1–6.

[35] Zellers R, Holtzman A, Rashkin H, et al. Defending against neural fake news[J]. Advances in neural information processing systems, 2019, 32.

[36] Fagni T, Falchi F, Gambini M, et al. TweepFake: About detecting deepfake tweets[J]. Plos one, 2021, 16(5): e0251415.

[37] gpt-2-output-dataset.2021.https://github.com/openai/gpt-2-output-dataset

[38] Uchendu A, Ma Z, Le T, et al. Turingbench: A benchmark environment for turing test in the age of neural text generation[C]// In Findings of the Association for Computational Linguistics: EMNLP 2021.2021: 2001–2016.

[39] He X, Shen X, Chen Z, et al. Mgtbench: Benchmarking machine-generated text detection[EB/OL].(2023-06-09).https://doi.org/10.48550/arXiv.2303.14822

[40] Liu Y, Zhang Z, Zhang W, et al. ArguGPT: evaluating, understanding and identifyingargumentative essays generated by GPT models[EB/OL].(2023-09-23).https://doi.org/10.48550/arXiv.2304.07666

[41] Li Y, Li Q, Cui L, et al. Deepfake Text Detection in the Wild[EB/OL].(2023-09-23).https://doi.org/10.48550/arXiv.2305.13242

[42] Wang Y, Mansurov J, Ivanov P, et al. M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection[EB/OL].(2023-05-24).https://doi.org/10.48550/arXiv.2305.14902

[43] Liu Z, Yao Z, Li F, et al. Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT[EB/OL].(2023-06-07).https://doi.org/10.48550/arXiv.2306.05524

[44] Mosca E, Abdalla M H I, Basso P, et al. Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era[C]//Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). 2023: 190-207.

[45] 范志武.基于深度金字塔卷积神经网络的 ChatGPT 生成文本检测方法[J/OL].数据分析与知识发现:1-14[2024-01-15].
[46] 邓胜利,汪璠,王浩伟.在线社区中人工智能生成内容的识别方法研究[J/OL].图书情报知识:1-11[2024-01-15].
[47] Fan A, Jernite Y, Perez E, et al. ELI5: Long form question answering[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.2019:3558–3567.
[48] Kočiský T, Schwarz J, Blunsom P, et al. The narrativeqa reading comprehension challenge[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 317-328.
[49] Shen Y, Rong W, Jiang N, et al. Word embedding based correlation model for question/answer matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2017: 31(1).
[50] He C, Deng Y, He L, et al. Engage Wider Audience or Facilitate Quality Answers? aMixed-methods Analysis of Questioning Strategies for Research Sensemaking on a Community Q&A Site[EB/OL].(2023-11-18). https://doi.org/10.48550/arXiv.2311.10975
[51] Jin Q, Dhingra B, Liu Z, et al. Pubmedqa: A dataset for biomedical research question answering[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),2019:2567–2577.
[52] Narayan S, Cohen S B, Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018:1797–1807.
[53] Li J, Sun M, Zhang X. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006: 545-552.
[54] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[J]. Advances in neural information processing systems, 2015, 28.
[55] O'Connor S. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?[J]. Nurse Education in Practice, 2022, 66: 103537-103537.
[56] Kang D, Ammar W, Dalvi B, et al. A dataset of peer reviews (peerread): Collection, insights and nlp applications[EB/OL].(2018-04-25).https://doi.org/10.48550/arXiv.1804.09635
[57] 陈悦,王智琦,刘则渊等.预印本的学术影响力研究——以 arXiv 自存档论文为例[J].情报学报,2019,38(08):815-825.
[58] 王颖,张智雄,钱力等.ChinaXiv 预印本服务平台构建[J].数字图书馆论坛,2017

(10):20-25.
[59] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL].2019.https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[60] Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies[J]. JMIR public health and surveillance, 2021, 7(11): e30642.
[61] Asghar N. Yelp dataset challenge: Review rating prediction[EB/OL].(2016-05-17).https://doi.org/10.48550/arXiv.1605.05362
[62] Kirchenbauer J, Geiping J, Wen Y, et al. A watermark for large language models[EB/OL].(2023-06-06).. https://doi.org/10.48550/arXiv.2301.10226
[63] Kirchenbauer J, Geiping J, Wen Y, et al. On the Reliability of Watermarks for Large Language Models[EB/OL].(2023-06-30).https://doi.org/10.48550/arXiv.2306.04634
[64] Zhao X, Li L, Wang Y X. Distillation-resistant watermarking for model protection in nlp[C]//Findings of the Association for Computational Linguistics: EMNLP 2022,5044–5055.
[65] Liu A, Pan L, Hu X, et al. A Private Watermark for Large Language Models[EB/OL].(2023-08-02).https://doi.org/10.48550/arXiv.2307.16230
[66] Dugan L, Ippolito D, Kirubarajan A, et al. RoFT: A tool for evaluating human detection of machine-generated text[EB/OL].(2020-10-06).https://doi.org/10.48550/arXiv.2010.03070
[67] Corston-Oliver S, Gamon M, Brockett C. A machine learning approach to the automatic evaluation of machine translation[C]//Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001: 148-155.
[68] Kalinichenko L A, Korenkov V V, Shirikov V P, et al. Digital Libraries: Advanced methods and technologies, digital collections[J]. D-Lib Magazine, 2003, 9(1): 1082-9873.
[69] Arase Y, Zhou M. Machine translation detection from monolingual web-text[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1597-1607.
[70] Beresneva D. Computer-generated text detection using machine learning: A systematic review[C]//Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21. Springer International Publishing, 2016: 421-426.
[71] Heymann,Paul,Koutrika,et al.Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges[J].IEEE Internet Computin

g, 2007, 11(6):36-45.
[72] Akram A. An Empirical Study of AI Generated Text Detection Tools[EB/OL].(2023-09-27).https://doi.org/10.48550/arXiv.2310.01423
[73] Lavergne T, Urvoy T, Yvon F. Detecting Fake Twitter Accounts with using Artificial Neural Networks[J]. Artificial Intelligence Studies, 2018, 1(1).
[74] Mindner L, Schlippe T, Schaaff K. Classification of Human-and AI-Generated Texts: Investigating Features for ChatGPT[C]//International Conference on Artificial Intelligence in Education Technology. Singapore: Springer Nature Singapore, 2023: 152-170.
[75] Mitchell E, Lee Y, Khazatsky A, et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature[EB/OL].(2023-07-23). https://doi.org/10.48550/arXiv.2301.11305
[76] Deng Z, Gao H, Miao Y, et al. Efficient Detection of LLM-generated Texts with a Bayesian Surrogate Model[EB/OL].(2023-05-26).https://doi.org/10.48550/arXiv.2305.16617
[77] Bao G, Zhao Y, Teng Z, et al. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature [EB/OL].(2023-10-08).https://doi.org/10.48550/arXiv.2310.05130
[78] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.
[79] Shevlane T, Farquhar S, Garfinkel B, et al. Model evaluation for extreme risks[EB/OL].(2023-09-22).https://doi.org/10.48550/arXiv.2305.15324
[80] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[EB/OL].(2019-07-26). https://doi.org/10.48550/arXiv.1907.11692
[81] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J].Advances in neural information processing systems, 2019, 32.
[82] Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[EB/OL].(2019-02-22).https://doi.org/10.48550/arXiv.1804.07461
[83] Bakhtin A, Gross S, Ott M, et al. Real or fake? learning to discriminate machine from human generated text[EB/OL].(2019-11-25).https://doi.org/10.48550/arXiv.1906.03351
[84] Antoun W, Mouilleron V, Sagot B, et al. Towards a Robust Detection of LanguageModel Generated Text: Is ChatGPT that Easy to Detect?[EB/OL].(2023-06-09).https://doi.org/10.48550/arXiv.2306.05871
[85] Sarvazyan A M, González J Á, Rosso P, et al. Supervised Machine-Generated Text Detectors: Family and Scale Matters[C]//International Conference of the Cross-Language Evaluation

Forum for European Languages. Cham: Springer Nature Switzerland, 2023: 121-132.
[86] Bhattacharjee A, Liu H. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?[EB/OL].(2023-08-17).https://doi.org/10.48550/arXiv.2308.01284
[87] Yu X, Qi Y, Chen K, et al. GPT Paternity Test: GPT Generated Text Detection with GPT Genetic Inheritance[EB/OL].(2023-05-21).https://doi.org/10.48550/arXiv.2305.12519
[88] Nelson E C, Hanna G L, Hudziak J J, et al. Obsessive-compulsive scale of the child behavior checklist: specificity, sensitivity, and predictive power[J]. Pediatrics, 2001, 108(1): e14-e14.
[89] Nhu V H, Mohammadi A, Shahabi H, et al. Landslide susceptibility mapping using machine learning algorithms and remote sensing data in a tropical environment[J]. International journal of environmental research and public health, 2020, 17(14): 4933.
[90] Sun Y, Wang S, Feng S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL].(2021-12-23).https://doi.org/10.48550/arXiv.2112.12731
[91] Christopoulou F, Lampouras G, Gritta M, et al. Pangu-coder: Program synthesis with function-level language modeling[EB/OL].(2022-07-22).https://doi.org/10.48550/arXiv.2207.11280
[92] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications[EB/OL].(2022-02-10).https://doi.org/10.48550/arXiv.2201.08239
[93] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.
[94] Levine Y, Dalmedigos I, Ram O, et al. Standing on the shoulders of giant frozen language models[EB/OL].(2022-04-21).https://doi.org/10.48550/arXiv.2204.10019
[95] Chaka C. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools[J]. Journal of Applied Learning and Teaching, 2023, 6(2).
[96] Liang W, Yuksekgonul M, Mao Y, et al. GPT detectors are biased against non-native English writers[EB/OL].(2023-07-10).https://doi.org/10.48550/arXiv.2304.02819
[97] Bloom B S, Krathwohl D R. Taxonomy of educational objectives: The classification ofeducational goals. Book 1, Cognitive domain[M]. longman, 2020.
[98] Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies[C]//AAAI workshop on learning from imbalanced data sets. AAAI Press Menlo Park, 2000, 68: 10-15.
[99] Bellinger C, Sharma S, Japkowicz N. One-class versus binary clas

sification: Which andwhen?[C]//2012 11th international conference on machine learning and applications. IEEE, 2012, 2: 102-106.
[100] Mitrović S, Andreoletti D, Ayoub O. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text[EB/OL].(2023-01-30).https://doi.org/10.48550/arXiv.2301.13852
[101] 王一博,郭鑫,刘智锋等.AI 生成与学者撰写中文论文摘要的检测与差异性比较研究[J].情报杂志,2023,42(09):127-134.
[102] Yang L, Jiang F, Li H. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text[EB/OL].(2023-07-21).https://doi.org/10.48550/arXiv.2307.11380
[103] Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence [J]. Electronic Markets,2021, 31: 447-464.
[104] 杨志航.算法透明实现的另一种可能：可解释人工智能[J/OL].行政法学研究:1-11[2024-01-11].
[105] Kowalczyk P, Röder M, Dürr A, et al. Detecting and Understanding Textual Deepfakes in Online Reviews[C]// Hawaii International Conference on System Sciences,2022.
[106] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey[EB/OL].(2023-09-19).https://doi.org/10.48550/arXiv.2309.07864
[107] Uchendu A, Lee J, Shen H, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts?[C]//Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2023, 11(1): 163-174.

作者贡献说明/Author contributions：

作者 1:王伟正：提出选题，收集数据与处理，撰写论文

作者 2:乔鸿：确定研究框架、论文修改

作者 3:李肖俊：撰写论文、修改论文

作者 4:王静静：修改论文

Research Progress of AIGC Recognition Enabled by Natural Language Processing Technology

Wang WeiZheng[1], Qiao Hong[2] , Li XiaoJun[3,4] and Wang JingJing[5]
(1. Library of Shandong Normal University, Jinan, 250358 ; 2.Business School of Shandong Normal University, Jinan 250358;3.Digital Humanities Research Center, Qilu University of Technology (Shandong Academy of Sciences),Jinan, 250014;4.Institute of Information, Qilu University of Technology (Shandong Academy of Sciences),Jinan, 250014;5.School of Journalism and Communication, Shandong University,Jinan, 250100)

Abstract: **[Purpose/Significance]**With the rapid ascent of large language models, AIGC have become ubiquitous in our daily lives. In order to mitigate potential misuse of AIGC, and to address issues such as the proliferation of false information, academic misconduct, and deceptive commentary, it is imperative to consolidate and forecast advancements in natural language processing techniques aimed at empowering AIGC discernment. ［**Method/Process**］ Firstly, it is essential to clarify that AIGC recognition constitutes a binary classification problem, with the aim of discerning whether a given piece of content is generated by artificial intelligence. Subsequently, employing a systematic review methodology, we have delineated the principal research outcomes in the domain of AIGC recognition.［**Result/Conclusion**］ The research identifies the critical significance of comprehensive and high-quality datasets in constructing classifiers for AIGC recognition. Simultaneously, it explores the limitations and developmental objectives of currently popular datasets, as well as potential datasets. Additionally, the paper analyzes paradigms of various classifiers, presents challenges across multiple domains such as multi-domain recognition tasks, cross-lingual recognition tasks, and data ambiguity issues. Finally, it summarizes the prospective development pathways for the future of AIGC recognition. This study aims to provide relevant researchers with a clear introduction and constructive suggestions for constructing more stable and efficient classifiers.
Keywords: AIGC; Machine-generated content detection; Black box test; white box test; Deep learning